Published in final edited form as:
J Phys Chem B. 2018 December 13; 122(49): 11295–11301. doi:10.1021/acs.jpcb.8b07206.

On the Natural Structure of Amino Acid Patterns in Families of Protein Sequences

Pablo Turjanski[*,†], **Diego U. Ferreiro**[*,‡]

[†]KAPOW, Departamento de Computación, Facultad de Ciencias Exactas y Naturales, UBA-CONICET-ICC, Buenos Aires, Argentina

[‡]Protein Physiology Lab, Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, UBA-CONICET-IQUIBICEN, Buenos Aires, Argentina

Abstract

All known terrestrial proteins are coded as continuous strings of ≈20 amino acids. The patterns formed by the repetitions of elements in groups of finite sequences describes the natural architectures of protein families. We present a method to search for patterns and groupings of patterns in protein sequences using a mathematically precise definition for “repetition”, an efficient algorithmic implementation and a robust scoring system with no adjustable parameters. We show that the sequence patterns can be well-separated into disjoint classes according to their recurrence in nested structures. The statistics of the occurrences of patterns indicate that short repetitions are sufficient to account for the differences between natural families and randomized groups of sequences by more than 10 standard deviations, while contiguous sequence patterns shorter than 5 residues are effectively random in their occurrences. A small subset of patterns is sufficient to account for a robust ”familiarity” definition between arbitrary sets of sequences.

Graphical Abstract

[*]**Corresponding Authors** (P.T.) pturjanski@dc.uba.ar., (D.U.F.) ferreiro@qb.fcen.uba.ar.

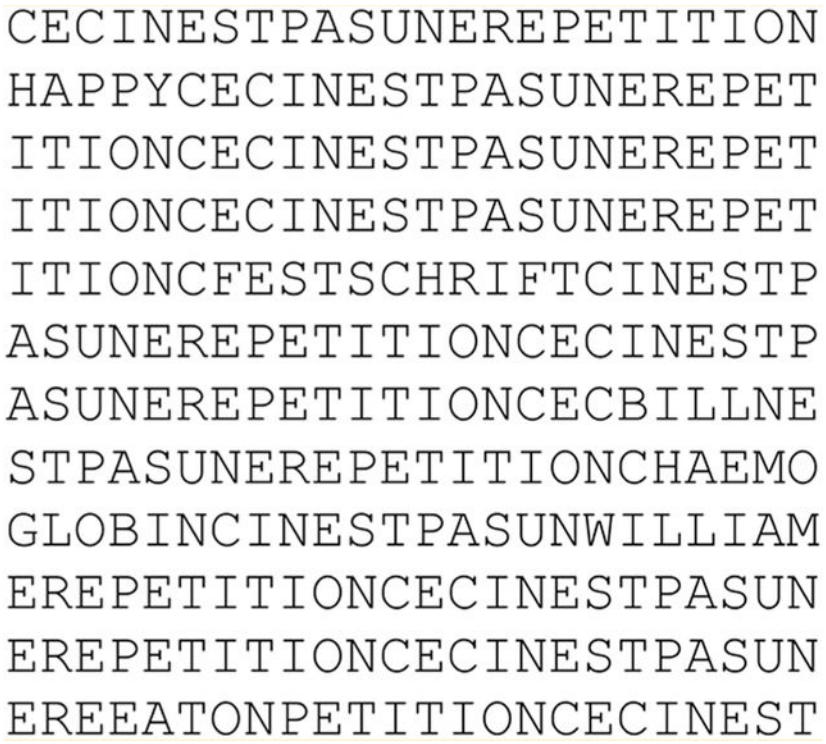
ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.8b07206.

Tables S1 and S2 with the protein families’ detailed data, pseudocode for the algorithmic implementation of MR search and match, and figures of the statistics of MR occurrences (PDF)

The authors declare no competing financial interest.



INTRODUCTION

> "See first, think later, then test. But always see first. Otherwise you will only see what you were expecting."
>
> Douglas Adams

Protein molecules can be described as finite linear strings of ≈20 amino acid types. It is still an intriguing fact that most natural amino acid strings appear indistinguishable from random by many statistical tests,[1] yet most of the random polypeptides synthesized do not behave as proteins do; they do not fold to specific structures nor do they *function* in a cellular context. Thus, the reduction in the description of proteins to linear strings of single amino acids misses a fundamental aspect to account for the, admittedly complex, biophysics of protein folding and function.[2,3] The search for "structural codes" in the analysis of protein sequences must consider the occurrence of correlations in the patterns of groups of amino acids, a task that gets combinatorially prohibitive to analyze exhaustively for all protein sequences.[4] Multiple heuristics designed to analyze correlations of the amino acid patterns in proteins have led to useful ways for approximating the grouping of sequences into families[5] and the structural ensembles of folded globules in these families[6] and even hint at the connection between folding thermodynamics[7] and the evolution of natural proteins.[8] Most of these methods require multiple sequences to be aligned each other through a common matrix in a so called multiple sequence alignment. Multiple sequence alignment is still a mathematical open problem, and thus the current computations of inferred alignments need to be tediously curated by human experts.[9]

To search for patterns and groupings of patterns in protein sequences, we use a mathematically rigorous definition of repetition and develop a method to characterize such repetitions. that uses no adjustable parameters. A maximal repetition (MR) is a well-defined exact match of a continuous block of amino acids that occurs two or more times in a single protein or in several proteins, while any of its extensions to longer sequences either to the N-terminus, the C-terminus, or both occurs fewer times. The search for maximal repetitions can be implemented with an algorithm whose computational complexity scaling is $O(n \log n)$, as n the size of the amino acid data set increases. This modest rate of growth in difficulty allows a very efficient exhaustive search.[10] The natural architectures of protein sequences can be analyzed by the occurrence of MR patterns. In previous work,[11] we introduced the concept and defined a continuos familiarity function that provides a fast quantification of the likelihood that any given amino acid string belongs to a given set of sequences. This *familiarity* function is computed from the search and match of maximal repetitions in sets of sequences. Here we show that the total of maximal repetition set can be well separated into disjoint classes according to their recurrence in nested structures. We analyze the statistics of maximal repetition classes in several natural protein families and in random strings, and find that only a small subset of the maximal repetitions is sufficient to account for a robust "familiarity" definition.

METHODS

Notation and Definitions.

Let there be an alphabet Σ, a finite set of symbols. We will consider linear sequences s of symbols in Σ of length $|s|$. We label the positions along the sequence s by counting from 1 to $|s|$. A string $s[i..j]$ denotes the sequence that starts at position i and ends at position j in s. If the proposition $1 \leq i \leq j \leq |s|$ is false, then $s[i..j]$ is equal to the empty sequence. We say u occurs in s if $u = s[i..j]$ for some i, j. A right extension of an occurrence $u = s[i..j]$ exists if $j < |s|$ and is $s[i..j + 1]$. A left extension of an occurrence $u = s[i..j]$ is said to exist if $i > 1$ and is denoted as $s[i - 1..j]$. A right context of an occurrence $u = s[i..j]$ is said to exist if $j < |s|$ and is denoted as $s[j + 1]$. A left context of an occurrence $u = s[i..j]$ exists if $i > 1$ and is denoted as $s[i - 1]$.

Definition 1.—(Gusfield[12]) A maximal repetition (MR) is a sequence that occurs more than once in s, and each of its extensions occurs fewer times.

We classify the different patterns of maximal repetitions into three disjoint categories: the first is super maximal repetition (SMR): which is a sequence that occurs more than once in s, while any of its extensions occurs only once. The next category is that of nested maximal repetition (NE), when all of the occurrences of the repetition are contained in a longer maximal repetition. Finally, the category of non-nested maximal repetition (NN), when at least one of the repetition occurrences are not contained in a longer maximal repetition and is not super maximal repetition. Formal details of these definitions have been described in a previous work.[13]

An illustration of the proposed classification of maximal repetitions (MR) is presented in Figure 1. The set of MRs of the string $s1 = cSMR1dSMR2eMRfSMR2gSMR1h$

(nonrepeating symbols are lower case) is {*MR, SMR, SMR*1, *SMR*2} (see Figure 1). Observe that *SMR*1 and *SMR*2 substrings are the longest maximal repetitions, occurring twice each. *SMR* and *MR* substrings are also maximal repetitions because they occurs four and five times in *s*1 respectively, and each of their extensions occurs fewer times. Note that *SM* is not a maximal repetitions because *SMR* (which is its unique possible repetitive right-extension) occur four times, violating the definition that any extension must occur fewer times. *SMR*1 and *SMR*2 are super maximal repetitions because their extensions occur only once. *SMR* is a nested maximal repetition since all of it occurrences are contained in *SMR*1 and *SMR*2. Finally, MR is classified as non-nested maximal repetition since, although 4 of its occurrences are contained in longer patterns (*SMR*1 and *SMR*2), there is a fifth occurrence that is not contained in any other longer repetition.

Definition 2.—Let S be a set of n sequences over the alphabet Σ, $S = \{s_1, s_2, \cdots s_n\}$. The set of maximal repetition in S is obtained by concatenation of all sequences in S, interleaved with different symbols $\$_1, \cdots, \$_{n-1}$ that are not in Σ. Thus, the set of maximal repetitions in S is the set of MR in $s_1\$_1 s_2\$_2 \ldots \$_{n-1} s_n$. If we work with sequences of characters, at the time of implementing this solution in a digital computer, there is an upper limit given by the necessary finite alphabet (Σ has only 256 symbols in the extended ASCII table) which restricts the amount of different $\$$ symbols we can use, and thus the number of sequences we can concatenate. To overcome this limitation, we implement a logically distinguished symbol (+), which we assign to it the unique property of being different from itself. That is, a non-Aristotelian $+ \neq +$. Taillefer et al.[13] proposed an algorithm that efficiently identifies and classifies MR from a sequence s into SMR, NE, and NN. We extended this algorithm in order to identify and classify MR originating from an arbitrary set of sequences (see the Supplementary Methods section in the Supporting Information).

RESULTS AND DISCUSSION

Occurrences of Maximal Repetition Patterns in Natural Proteins.

In order to analyze the structure of maximal repetition patterns in natural protein sequences we concentrate on 46 abundant protein families which have been curated. Each of the families contains between 924 and 38 342 nonredundant sequences, and comprise between 805 684 and 23 670 587 amino acids (Table S2), making a grand total of 696 114 strings and 434 447 858 amino acids. We analyze some families for which recurrent structural repetitions have already been annotated (as “repeat-proteins”[14]), and families for which no such repetitions have previously been reported (called ”globular proteins”). For each family the distribution of maximal repetitions was calculated and each MR classified as either SMR, NN or NE (see Figure 1). The relative populations of each maximal repetition class in each family are shown in Figure 2. Overall, the distribution of MR classes appears roughly constant between families: most of the MR are non-nested (NN), around 20% are nested (NE), and about 25% are true supermaximal repeats (SMR). This distribution holds irrespectively of the common classification of repeat vs globular protein family, indicating that the overall pattern of nesting of multiple repetitions is a general characteristic of all protein sequences. One clear exception is the Nebulin family, for which we identify an overabundance of nested repeats. It was previously reported that the repetitions found in this

family can be described as short repeats within longer repeats,[15] which we identify as nested occurrences (Figure 2).

To test whether the distributions of MR types is random or differs in some way characteristic for natural protein families, we constructed three control groups of sequences: *RandomAA* is a set of sequences drawn entirely by chance of 20 letters each with equal probabilities. *ScrambledAA* is an exhaustive permutation of the amino acids of one natural family, thus conserving the natural bias in the amino acid composition[16] and Heterogeneous is a set of natural sequences picked at random from all the families (see Supplementary Methods for details). All three control groups show a common distribution of abundance of maximal repetitions categories, with super maximal repetitions being the most prevalent and only a minority of nested repetitions (Figure 2). For the *RandomAA* and the *ScrambledAA* controls, it can be expected that most of the maximal repetitions will not be found in the nested category, as the nesting probability of maximal repetitions decreases exponentially in random strings,[17] and thus super maximal repetitions will prevail. However, the *Heterogeneous* picking of sequences from various families shows a similar distribution, hinting that the nesting patterns are properties emerging from the grouping of sequences, and are not found at the level of individual proteins.

The abundances of distinct types of maximal repetitions in the families could depend on the length of the MR set under scrutiny: shorter maximal repetitions are trivially more prevalent than longer ones in any string. Figure 3A shows that all of the repetitions of length 1 and 2 amino acids are nested in longer MR in all families. The fraction of nested repetitions drops to about 10% at length 5 and then grows to about 40% for lengths of few decades. The non-nested repeats are most prevalent at length of 4 to 5 amino acids and the super maximal repetitions are the most abundant when the longer MRs are considered. The relative abundance of each maximal repetitions class shows a complicated length dependence, that we find consistently in each family. It is expected that super maximal repetitions will have to be the most prevalent class at the longest lengths, as every nested or non-nested is ultimately contained within an a larger super maximal repetitions (Figure 1). The same analysis performed on the random set indeed shows that super maximal repetitions is the only class of maximal repetitions at lengths larger than 8 amino acids, with non-nested being the most prevalent at length 5 and nested absent above length 7 (Figure 3B, dashed lines). In contrast, when multiple sequences are grouped into an artificial control family, non-nested and nested maximal repetitions persist up to length 100 (Figure 3B, continuous lines).

If there is structure in the architecture of the repetitions in a finite string, it is expected that not all maximal repetitions will be equally abundant. We quantified the total number of different patterns in all the maximal repetitions classes in all families. As can be seen in Figure 3 C, all of the maximal repetitions of size 1 (20 single amino acids) are present in all families, and every occurrence is nested in longer MRs. From the millions of distinct maximal repetitions found in each family, most of them have lengths between 4 and 8 amino acids, being the most prevalent SMR and NN types of 4 to 6 amino acids. Notably, for MR longer than 10 residues, the distribution appears to follows a power law where $MR_{Total\,Number} \approx a * MR_{length}^{\gamma}$. The γ exponent is about −2.6 regardless of the maximal

repetition class. This value for γ is clearly not the case for the control randomized set, where there are no MRs larger than 12 amino acids and the γ exponent is about −10 (Figure 3 D, dashed lines). When multiple sequences are grouped into an artificial control family (Figure 3 D, continuous lines) γ is around −3.8. The γ exponent appears to be similar for each and every protein family (Figure S1).

To set the length scale for maximal repeat evaluation, we calculated the fraction of the possible strings that are present as MR in natural sequences. Every single and all possible pairs and triplets of amino acids can be found in the natural data set, and these are typically nested in longer MRs (Figure 4A). As the possible number of amino acid sequences grows with length as 20^N, the coverage of the sequence space precipitously drops and only a few of the possible MRs of length longer than 6 can be found. Both control sets show a slightly higher coverage of the sequence space than the natural families, and super maximal repetition and non-nested are found in larger proportions than the nested ones at length 5 amino acids (Figure 4B). Notably, both the random and the artificial family sets display equivalent coverage of the possible sequences at short sequence length, suggesting that natural proteins look effectively random at lengths shorter than 4 amino acids. The artificial grouping of sequences in the *Heterogeneous* control set explores the sequence space equally well as the random set, up to a length of 7 amino acids (Figure 4B).

Sequence Coverage and Familiarity.

In general, all of protein families that were analyzed display a similar distribution of maximal repeat patterns in terms of the lengths and of the repetitions MR class (Figure 3). Nevertheless, the specific sequences of the MR sets for each family can be very different, as they account for an almost insignificant proportion of all the possible amino acid strings for lengths than are larger than 6 residues (Figure 4). To evaluate how for the distinct MR sets can go in accounting for the occurrence of specific patterns in natural protein sequences, we developed two continuous evaluation functions which we call coverage and familiarity.[11] Briefly, the function *coverage*: $\Sigma^* \times \mathcal{P}(\Sigma^*) \to \mathbb{Q}$ is defined for any sequence s and any set of sequences R through the relation:

$$coverage(s, R) = \frac{\#\{j : \exists\ i \in \mathbb{N}, \exists\ r \in R, s[i..i + |r| - 1] = r\}}{|s|} \quad (1)$$

We see that *coverage*(s,R) is a rational number between 0 and 1. Figure 5A shows *coverage*(s,R) evaluated on the string of the natural protein IκBα of *H. sapiens* with distinct MR subsets. The sequence can be covered fully with most short MRs, and *coverage*(s, R) is larger than 0.9 for all MR subsets originating from the ANK family up to minimum pattern length 10. In contrast, *coverage*(s, R) for IκBα drops to zero at minimum pattern 7 amino acids when the MR sets are originated from the ABCTran family. This result is not surprising as IκB proteins have been annotated as containing ankyrin repeat regions (grouped in the ANK family), and no ABCtran family regions.[18]

The *familiarity* function $\Sigma^* \times \Sigma^* \to \mathbb{Q}$ measures how much of a sequence is covered by a set of MRs. For any two sequences s and t, the familiarity of the pair is given by

$$familiarity\,(s,\ t) = \frac{coverage(s, \mathscr{M}(t,0)) + coverage(s, \mathscr{M}(t, |s|))}{2} + \sum_{i=1}^{|s|-1} coverage(s, \mathscr{M}(t,i)) \quad (2)$$

where $\mathscr{M}(t,n)$ denotes the set of MRs from t of lengths greater than or equal to n. $\mathscr{M}(t,0)$, by definition, returns all the blocks of the sequence t. Computing *familiarity* requires finding the values of the $coverage\,(s, \mathscr{M}(t,i))$ for each i in $[0..|s|]$, which we find is enough to limit to [0..10]. This allows for a robust comparison of sequences s of different lengths.[11] *familiarity*(*s,t*) is thus a rational number between 0 (when not a single part of *s* can be covered by MRs of *t*, only possible for disjoint alphabets) and 10 (when the whole *s* can be covered with MRs of *t*). As the second argument *t* of the familiarity function we denote the concatenation of all the sequences of a group separated by the distinguished symbol (+) (Definition 2).

Figure 5B shows the values for the *familiarity* evaluated over 10 natural test sequences from the Ankyrin family with different sets of maximal repeats. All of these sequences score over 6 in *familiarity* when the MR set *t* originates from the ANK family, as expected, as these sequences are annotated to have ankyrin regions (Table S1). However, the *familiarity* is around 6 when *t* is constructed from the control sequences. These values of *familiarity* originate from the common underlying structure of the patterns of both natural and random sequences up to a length of 5 amino acids (*vide supra*). Both nested and non-nested subsets account for these distinctions and the values spread for the super maximal repeats subsets (Figure 5B). Thus, these sequences can be similarly well described with the structural patterns of the nested and non-nested subsets.

Natural protein sequences often encode distinct functional *domains*. These domains are usually reflected as common structural patterns that persist over evolutionary times, and may be sometimes artificially decoupled along the amino acid strings.[19] These biological lumping must be related to the maximal repeat patterns found in the sequence descriptions. To investigate how the maximal repeat patterns of the natural protein families differ from those of random strings, we computed *familiarity*(*s,t*) for 10 test sequences that have been annotated to belong to each of the 46 families under scrutiny, but that are not present in *t*. To evaluate the maximal repeat subsets on common grounds, we compare the Z-scores of the *familiarity*(*s,t*) distributions of the test sequences with respect to those for the random sets Z-score = (*familiarity*(*s, t*) − *AVG*(*familiarity*(*ScrambledAA_test,t*)))/ *STD*(*familiarity*(*ScrambledAA_test,t*)) where *ScrambledAA_test* are the ten test synthetic proteins from *ScrambledAA* control family (Figure 6). Both the non-nested and nested subsets are excellent at distinguishing the test protein sets from the random sequences, as the mean Z-score is above 10 for most families. In most cases, both nested and non-nested are as good as the whole maximal repeat set. Some families show consistently larger Z-scores (Nebulin and PPLlike), with a somehow larger mean values for the families grouped as repeat-proteins (Figure 6). For all families, the Z-scores of the super maximal repeat subsets is lower, indicating that the test sequences cannot be well explained with these patterns. To

analyze if combinations of the maximal repeat subsets significantly alter the *familiarity*(*s,t*) scoring, we constructed all pair-combinations of non-nested, nested, and super maximal repeat in *t* and found that none of these significantly perturb the results (Figure 6).

To investigate the way maximal repeat patterns between the families overlap, we computed *familiarity*(*s,t*) for 460 test sequences, 10 *s* for each of the 46 *t* families under scrutiny (Table S1). In Figure 7 the unique Uniprot sequence entries have been ordered according to the presence of at least one PFAM domain. The strong diagonal of high *familiarity*(*s,t*)values thus reflects that the PFAM grouping is consistent with the definitions, the computation and the scorings we propose. Some families display consistently low values of *familiarity* toward all sequences (Nebulin), and some consistently higher values (HelicaseC). It is also apparent that some families are clearly related (LdlReceptA and LdlReceptB), even when their historical naming differs for the two families (ARM and HEAT). In some cases, a given sequence displays significant *familiarity* to more than one single family, hinting at the presence of multiple biological domains in that sequence. In some other cases, groups of test sequences only display *familiarity* toward one single family (TSP). The multiple ”bands” that are apparent in the representation of the data in Figure 7 are probably not random but a manifestation of some deeper structure in the original data, which deserves further investigation but is out of the scope of the present report. We note that the results are robust to the choice of the subsets of MRs that were used to compare the sequences and combinations thereof (Figure S2).

DISCUSSION

For *genetic information* to be a meaningful modern concept, natural protein sequences cannot be just random strings of amino acids.[20–22] Spontaneous, fast and robust folding of polypeptide chains is the organic way in which structural patterns emerge from amino acid sequences in certain environments.[23] The search for underlying *folding codes* to de/construct the folding energy landscapes involves the realization that at some level natural sequences are fundamentally distinguishable from random strings,[24] yet the actual correspondences are clearly complex, as they involve a myriad of small, nonlocal, interactions. Effective ways of reverse engineering folding have been achieved at different levels of description using clever heuristics.[25–27] This fact indicates that it is possible to deconvolute the physical phenomenon of biological molecules without directly studying them at the fundamental quantum mechanical level. Searching for the historical footprints in the extant sequences has led to useful approximations for the exploration of the energy landscapes of structural[7] as well as the sequence spaces.[28]

All known terrestrial proteins can be described as linear repetitions of amino acids. We searched for patterns and groupings of patterns in natural protein sequences using a mathematically rigorous definition for the concept of “repetition” (MR, Definition 1), an efficient algorithmic implementation (Supporting Information, code2) and a robust scoring system for repetitions that did not require introducing any adjustable parameters.[11] We propose that the maximal repeat set computed for a group of sequences can be well-separated into disjoint classes (Figure 1). Each maximal repeat is either supermaximal (SMR), nested (NE), or non-nested (NN) according to the patterns of occurrence in a given

set of sequences. The relative populations of each MR class in natural protein families are similar in all families but population of repetitions class are clearly distinguishable from the randomized control groups of sequences (Figure 2). When natural sequences are randomly grouped into artificial families, one finds similar total maximal repeat fractions as for controls in which the sequences themselves are randomized. This observation indicates that the nesting patterns of the repetitions are the main objects underlying the distributions found in the natural sets. Indeed, the frequency of occurrence of repeats shorter than 5 amino acids is equivalent in both the natural and the artificial randomly constructed sets of sequences (Figure 3), covering the sequence space as expected for the exhaustive exploration of patters in random sets of similar, finite, size (Figure 4).[29] As the sequence space grows exponentially with string length, almost none of the full sequence space can be covered by repetitions larger than 5 amino acids. Nevertheless, the occurrence of patterns of repetitions in natural sequences is clearly not random in any family and most of the changes in the distributions of maximal repeats occurs between 5 and 10 amino acids (Figure 3). Perhaps it is not a coincidence that regular secondary structure elements first made their appearance at this length scale.[30] This length also signals the critical size window at which good structure prediction heuristics work and where *foldons* have been predicted to emerge.[31] The patterns of repetitions that are larger than 10 amino acids can be crudely described by a power law distribution for all natural protein families. The γ exponent is about −2.6 for all families and maximal repeat subsets (Figure 3 and Figure S1). This apparent scale invariant distribution of structure in natural proteins has previously been hold at the tertiary level, and has been alleged to the fractal geometries of natural folds.[32–35] Comparable exponents were also reported for the distribution of voids in the interior of protein *swiss-cheese* globules.[36] Thus, there is an apparent common structure of amino acid patterns larger than 10 residues that can be detected in the primary structure of protein families and at the tertiary level of individual folded proteins.

Searching for MRs and matching them to sequences can be efficiently used to characterize the structure of any sequence s with respect to a set of sequences t, by computing the *familiarity*(s,t) function.[11] Both the nested and non-nested subsets of maximal repeats are good descriptors of *familiarity*(s,t) as is the whole maximal repeat set (Figure 6). The overall patterns of repetitions shorter than 10 residues is enough to account for the differences between natural and random sequences by more than 10 standard deviations (Figure 6). The scoring we put forward is robust to the combinations of maximal repeat subsets, and the exhaustive search of the SMRUNN subset can be implemented with an algorithm whose computational complexity is $O(n)$ (Suppl. code2).

Natural protein sequences encode *functional domains* of finite size.[19] The biological accretion of *functional information* can be expected to be detectable at the lengths scales at which proteins differ from the occurrence of patterns in random strings.[37] Computing the *familiarity*(s,t) for groups of existing proteins indeed reveals exciting patterns of common structure that are discernible at the length scales of 5−10 amino acids (Figure 7). The *familiarity*(s,t) distributions are robust to the MR subsets used and indicate that PFAM grouping is consistent with the definitions, the computation and the scorings we propose (Figure S2). In some cases, a given sequence displays significant *familiarity* to more than one family, hinting to the presence of multiple biological domains. In some other cases,

groups of test sequences score consistent *familiarity* toward one single family. Presumably evolutionary relationships between groups of sequences can also be detected as groups that score consonant between PFAM families (Figure 7). Since *familiarity*(*s*,*t*) is a well-defined continuous function and the maximal repeat search can be exhaustively carried out with existing computers, *familiarity* analysis could be used as a general tool to explore the biological relationships between arbitrary groups of sequences. Developing appropriate metrics in the sequence space[38] together with efficient search strategies can hint at the length scales at which the *natural coding* of biological information originates.[39–42]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Pablo Rago and Vero Becher for their support and insightful discussions, Peter G. Wolynes for passionately sharing grammatical knowledge, and Bill Eaton for constant inspiration. This work was supported by the Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET), the Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), Ecos-Sud. NASA Astrobiology Grant Number 80NSSC18M0093 to the ENIGMA team is acknowledged.

REFERENCES

(1). Weiss O; Jiménez-Montaño MA; Herzel H. Information Content of Protein Sequences. J. Theor. Biol 2000, 206, 379–386. [PubMed: 10988023]

(2). Wolynes PG; Eaton WA; Fersht AR Chemical physics of protein folding. Proc. Natl. Acad. Sci. U. S. A 2012, 109, 17770–1. [PubMed: 23112193]

(3). Eaton WA; Wolynes PG Theory, simulations, and experiments show that proteins fold by multiple pathways. Proc. Natl. Acad. Sci. U. S. A 2017, 114, E9759–E9760. [PubMed: 29087352]

(4). Dryden DTF; Thomson AR; White JH How much of protein sequence space has been explored by life on Earth? J. R. Soc., Interface 2008, 5, 953–6. [PubMed: 18426772]

(5). Doolittle RF The roots of bioinformatics in protein evolution. PLoS Comput. Biol 2010, 6, e1000875. [PubMed: 20686682]

(6). Morcos F; Pagnani A; Lunt B; Bertolino A; Marks DS; Sander C; Zecchina R; Onuchic JN; Hwa T; Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U. S. A 2011, 108, E1293–301. [PubMed: 22106262]

(7). Schafer NP; Kim BL; Zheng W; Wolynes PG Learning To Fold Proteins Using Energy Landscape Theory. Isr. J. Chem 2014, 54, 1311–1337. [PubMed: 25308991]

(8). Morcos F; Schafer NP; Cheng RR; Onuchic JN; Wolynes PG Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc. Natl. Acad. Sci. U. S. A 2014, 111, 12408–13. [PubMed: 25114242]

(9). Dickson RJ; Wahl LM; Fernandes AD; Gloor GB Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. PLoS One 2010, 5, e11082. [PubMed: 20596526]

(10). Becher V; Deymonnaz A; Heiber P. Efficient computation of all perfect repeats in genomic sequences of up to half a gigabyte, with a case study on the human genome. Bioinformatics 2009, 25, 1746–1753. [PubMed: 19451169]

(11). Turjanski P; Parra RG; Espada R; Becher V; Ferreiro DU Protein Repeats from First Principles. Sci. Rep 2016, 6, 23959. [PubMed: 27044676]

(12). Gusfield D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology; Cambridge University Press: 1997.

(13). Taillefer E; Miller J. Exhaustive computation of exact duplications via super and non-nested local maximal repeats. J. Bioinf. Comput. Biol 2014, 12, 1350018.

(14). Di Domenico T; Potenza E; Walsh I; Gonzalo Parra R; Giollo M; Minervini G; Piovesan D; Ihsan A; Ferrari C; Kajava AV; et al. RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Res. 2014, 42, D352–D357. [PubMed: 24311564]

(15). rklund AK; Light S; Sagit R; Elofsson A. Nebulin: a study of protein repeat evolution. J. Mol. Biol 2010, 402, 38–51. [PubMed: 20643138]

(16). Krick T; Verstraete N; Alonso LG; Shub DA; Ferreiro DU; Shub M; Sanchez IE Amino Acid metabolism conflicts with protein diversity. Mol. Biol. Evol 2014, 31, 2905–12. [PubMed: 25086000]

(17). Crochemore M; Rytter W. Jewels of Stringology; World Scientific: 2002.

(18). Trelle MB; Ramsey KM; Lee TC; Zheng W; Lamboy J; Wolynes PG; Deniz A; Komives EA Binding of NFκB Appears to Twist the Ankyrin Repeat Domain of IκBα. Biophys. J 2016, 110, 887–95. [PubMed: 26910425]

(19). Espada R; Parra R; Sippl M; Mora T; Walczak A; Ferreiro D. Repeat proteins challenge the concept of structural domains. Biochem. Soc. Trans 2015, 43, 844–849. [PubMed: 26517892]

(20). Smith JM The Concept of Information in Biology. Philosophy of Science 2000, 67, 177–194.

(21). Adami C. Information theory in molecular biology. Physics of Life Reviews 2004, 1, 3–22.

(22). Godfrey-Smith P. Information in biology. The Cambridge Companion to the Philosophy of Biology 2007, 103–119.

(23). Ferreiro DU; Komives EA; Wolynes PG Frustration in biomolecules. Q. Rev. Biophys 2014, 47, 285–363. [PubMed: 25225856]

(24). Bryngelson JD; Wolynes PG Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. U. S. A 1987, 84, 7524–8. [PubMed: 3478708]

(25). Muñoz V; Eaton WA A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc. Natl. Acad. Sci. U. S. A 1999, 96, 11311–6. [PubMed: 10500173]

(26). Robustelli P; Piana S; Shaw DE Developing a molecular dynamics force field for both folded and disordered protein states. Proc. Natl. Acad. Sci. U. S. A 2018, 115, E4758–E4766. [PubMed: 29735687]

(27). Leaver-Fay A; Tyka M; Lewis SM; Lange OF; Thompson J; Jacak R; Kaufman K; Renfrew PD; Smith CA; Sheffler W; et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011, 487, 545–74. [PubMed: 21187238]

(28). Bornberg-Bauer E; Chan HS Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc. Natl. Acad. Sci. U. S. A 1999, 96, 10689–94. [PubMed: 10485887]

(29). Lavelle DT; Pearson WR Globally, unrelated protein sequences appear random. Bioinformatics 2010, 26, 310–8. [PubMed: 19948773]

(30). Parra RG; Espada R; Sanchez IE; Sippl MJ; Ferreiro D U. Detecting repetitions and periodicities in proteins by tiling the structural space. J. Phys. Chem. B 2013, 117, 12887–97. [PubMed: 23758291]

(31). Panchenko AR; Luthey-Schulten Z; Cole R; Wolynes PG The foldon universe: a survey of structural similarity and selfrecognition of independently folding units. J. Mol. Biol 1997, 272, 95–105. [PubMed: 9299340]

(32). Stapleton HJ; Allen JP; Flynn CP; Stinson DG; Kurtz SR Fractal Form of Proteins. Phys. Rev. Lett 1980, 45, 1456–1459.

(33). Lewis M; Rees DC Fractal surfaces of proteins. Science 1985, 230, 1163–5. [PubMed: 4071040]

(34). Reuveni S; Granek R; Klafter J. Anomalies in the vibrational dynamics of proteins are a consequence of fractal-like structure. Proc. Natl. Acad. Sci. U. S. A 2010, 107, 13696–700. [PubMed: 20639464]

(35). Kornev AP Self-organization, entropy and allostery. Biochem. Soc. Trans 2018, 46, 587–597. [PubMed: 29678954]

(36). Chowdary PD; Gruebele M. Molecules: what kind of a bag of atoms? J. Phys. Chem. A 2009, 113, 13139−43. [PubMed: 19588898]

(37). Dayhoff MO The origin and evolution of protein superfamilies. Fed Proc. 1976, 35, 2132–2138. [PubMed: 181273]

(38). Schwende I; Pham TD Pattern recognition and probabilistic measures in alignment-free sequence analysis. Briefings Bioinf. 2014, 15, 354–368.

(39). Rodríguez PE Dogma periférico: ?de qué mensaje me están hablando? Química Viva 2015, 14, 1–10.

(40). Kirschner M; Gerhart J; Mitchison T. Molecular ”vitalism. Cell 2000, 100, 79–88. [PubMed: 10647933]

(41). Ferreiro DU; Komives EA; Wolynes PG Frustration, function and folding. Curr. Opin. Struct. Biol 2018, 48, 68–73. [PubMed: 29101782]

(42). Adams D. The Salmon of Doubt: Hitchhiking the Galaxy One Last Time; William Heinemann Ltd: 2002.

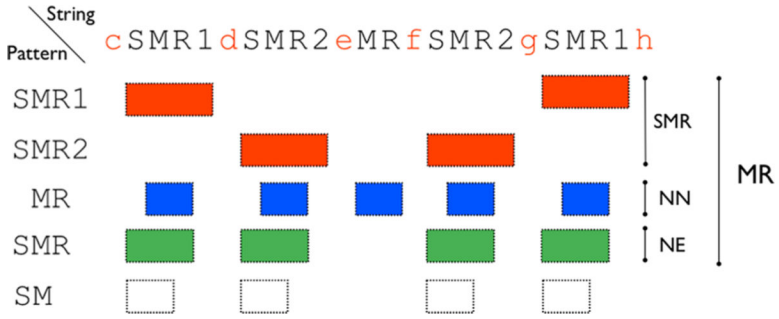


Figure 1.

Maximal repetitions (MR) computed for the input string shown on top (nonrepeating symbols are in red lower case). MR patterns are classified in disjoint subgroups according to their patterns of occurrence as super maximal repetition (SMR), non-nested maximal repetition (NN) and nested maximal repetition (NE).

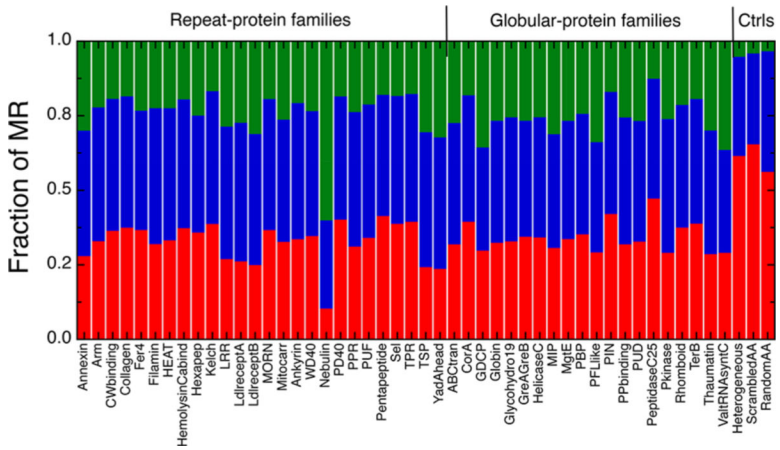


Figure 2.

Fraction of MR in each class for all protein families analyzed. Nested maximal repeats (NE, green), non-nested maximal repeats (NN, blue) and supermaximal repeats (SMR, red) were computed in each natural family of either globular or repeat-protein classes. Ctrls indicates the three control groups of artificial families (see main text).

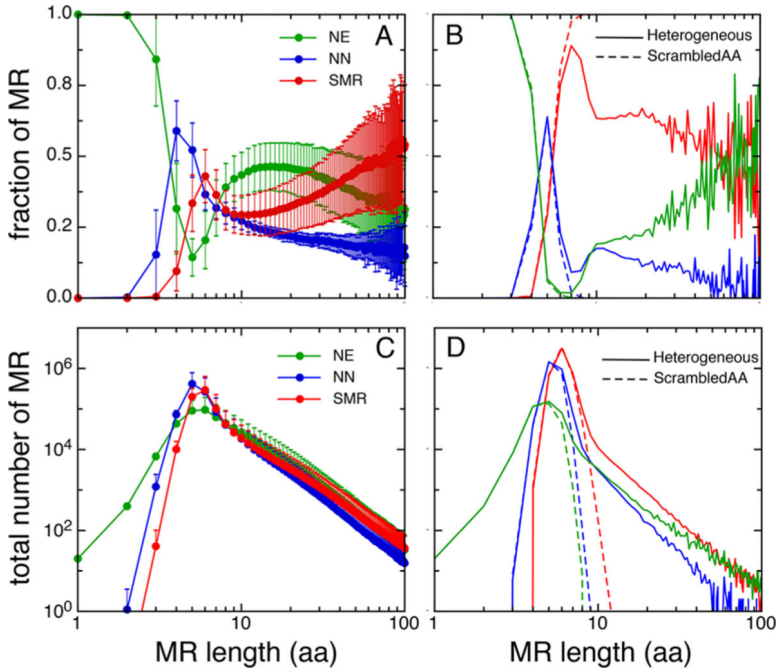


Figure 3.

Distribution of MR patterns in each subset. The MRs were calculated for each family and the mean and standard deviation of all families is shown. The relative abundance at different lengths is shown in part A and the total abundance in part C. Equivalent calculations of the control families are shown in parts B and D.

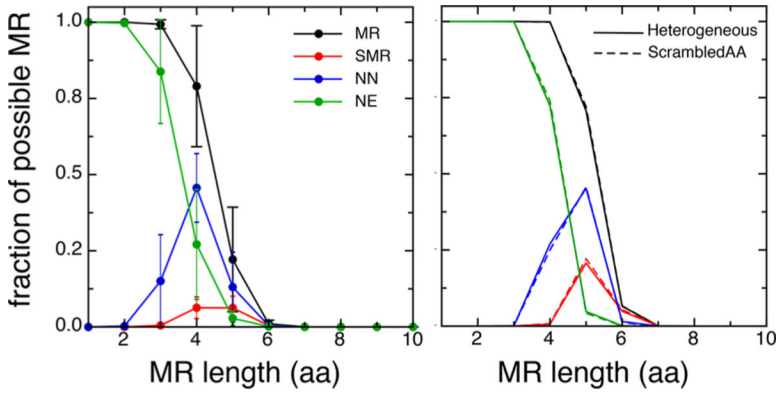


Figure 4.

Coverage of the sequence space. The MRs were calculated for each family and the fraction of the total possible patterns is shown for each subset. The mean and standard deviation of all families is shown in part A and for control groups in part B.

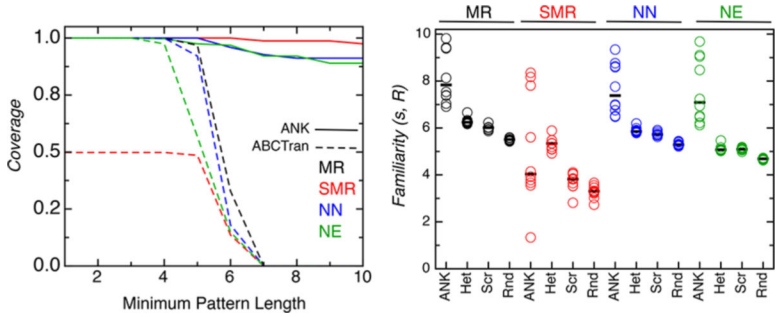


Figure 5.

Evaluating *coverage*(*s*,*R*) and *familiarity*(*s*,*t*) functions on natural proteins. (A) Values for *coverage* for the sequence *s* corresponding to the natural protein I*κ*B*α* of *Homo sapiens*, with distinct MR subsets *t*. (B) Values for *familiarity* for 10 natural protein sequences *s* annotated to have Ankyrin repeats, evaluated with different MR subsets from the ANK family and three control groups.

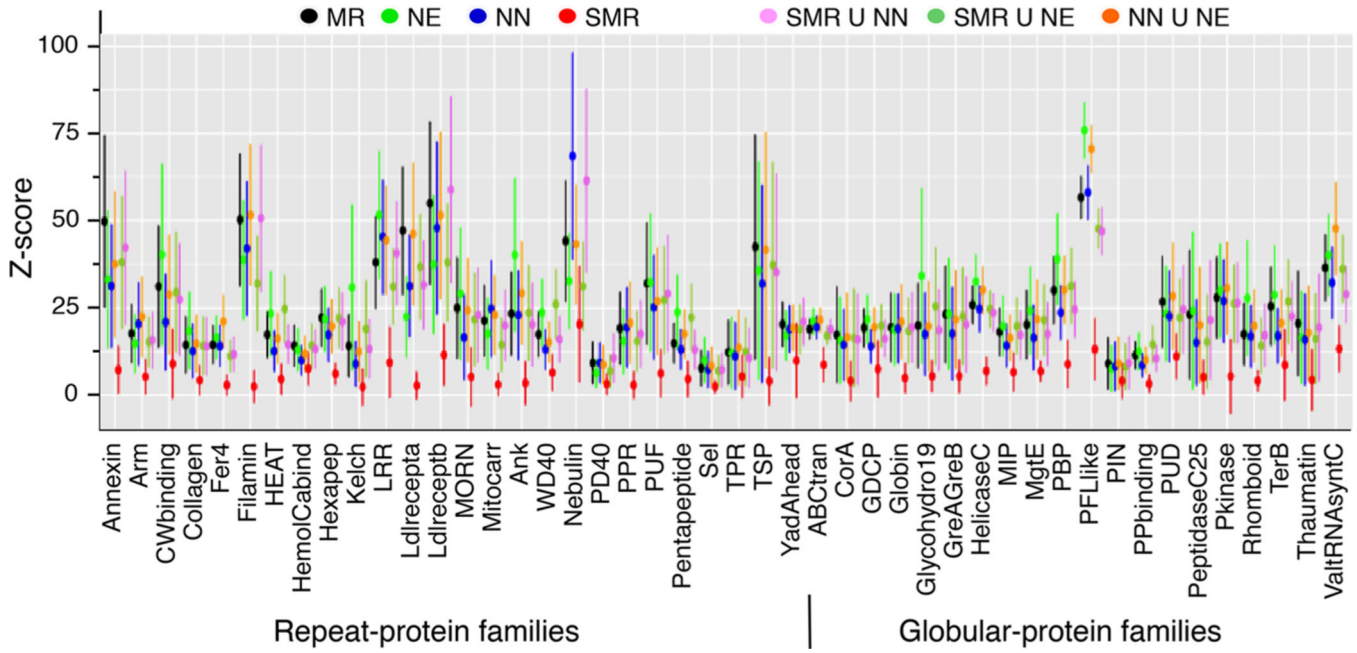


Figure 6.

Evaluating *familiarity*(*s,t*) for natural sequences with random MR subsets. *Familiarity*(*s,t*) was computed for ten *s* sequences of each natural protein family and the set *t* of ScrambledAA control group. The mean and standard deviation of the *Z*-score distributions are shown, computed with the NN, NE, and SMR subsets and all the pair-unions of these, SMRUNN, SMRUNE, and NNUNE.

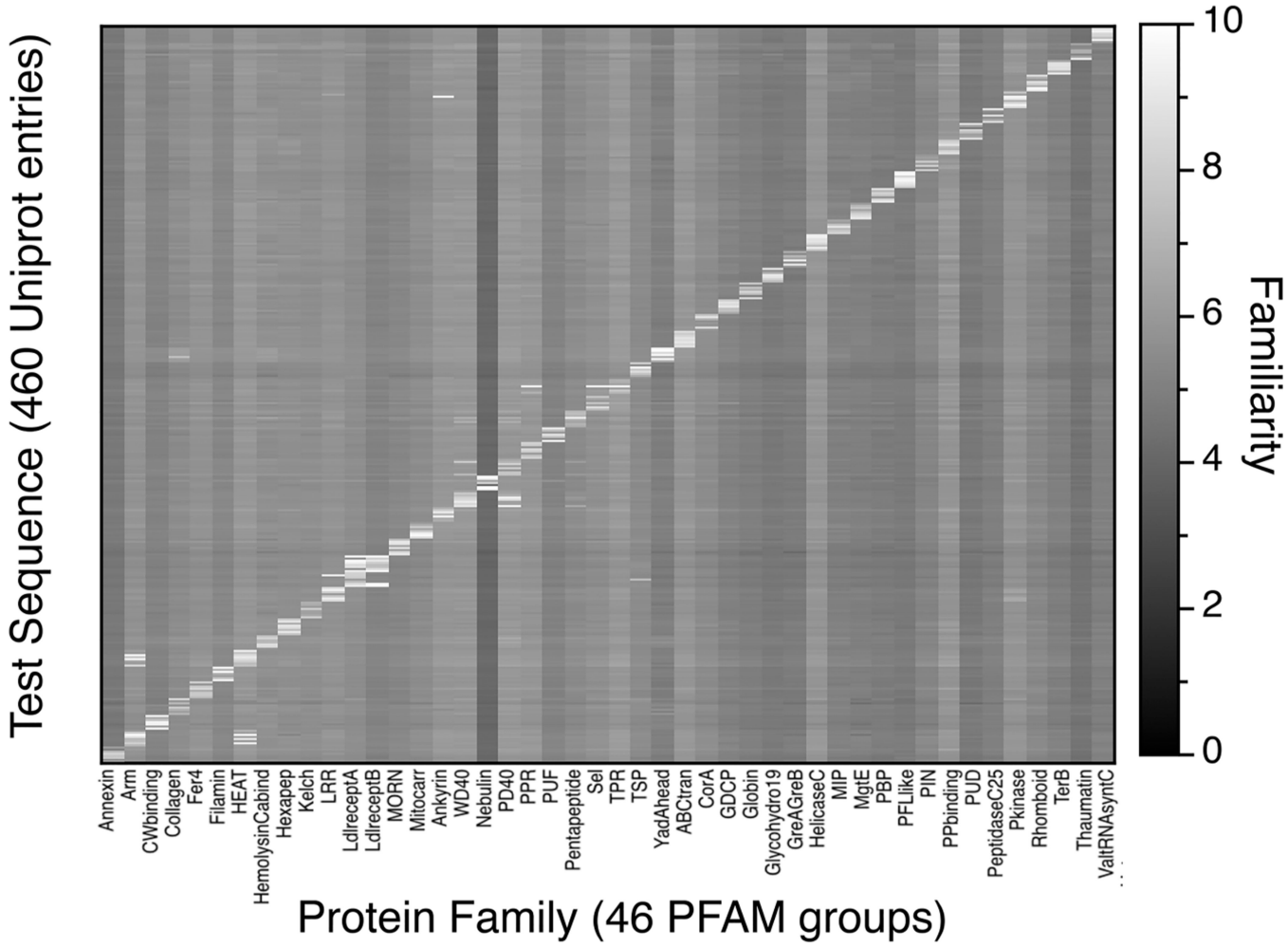


Figure 7.

Evaluating *familiarity*(*s,t*) for natural sequences and natural families. *Familiarity*(*s,t*) was computed for 460 test sequences *s* and 46 *t* families with the SMRUNN subsets of MR. The unique Uniprot entries are ordered according to the presence of at least one PFAM domain.